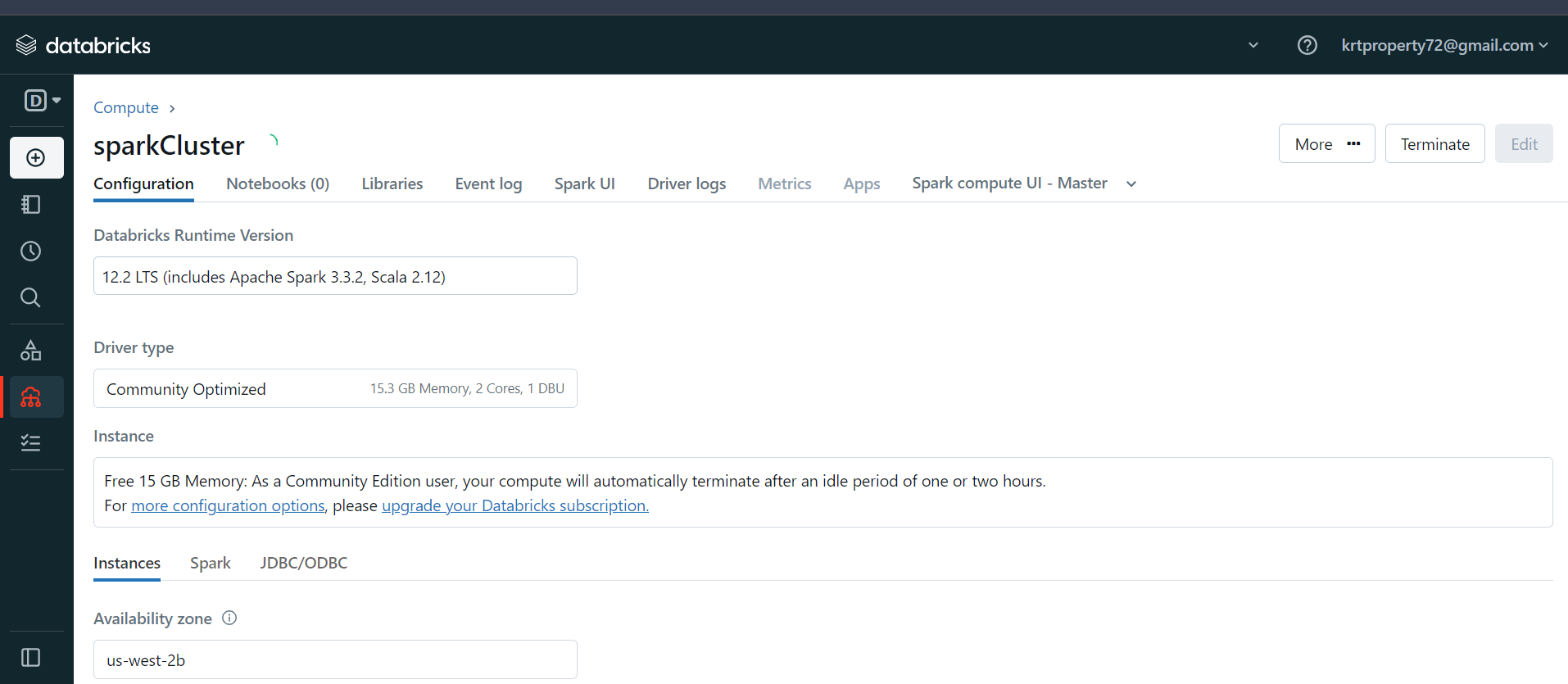
**Data Analysis, Visualization and Prediction using Pyspark in Databricks**

**Step – 1 :** Create Databricks account

A screenshot of a computer

Description automatically generated

Step – 2 : Create a compute engine which will initialize spark cluster



Step – 3 : Now, download and extract the dataset file

A screenshot of a computer

Description automatically generated

Step – 4 : Now, upload the dataset into spark catalog

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step – 5 : Click on Create Table in Notebook, then it will open new notebook

A screenshot of a computer

Description automatically generated

Step – 6 : Create a pyspark dataframe and display it

A screenshot of a computer

Description automatically generated

Step – 7 : preprocess the data, remove all the unwanted columns

A screenshot of a computer

Description automatically generated

Step – 8 : Check columns, whether they are updated or not

A white background with black text

Description automatically generated

Step – 9 : Now, check the datatypes of all columns

A screenshot of a computer

Description automatically generated

Step – 10 : Rename the columns and replace space with ‘\_’

A screenshot of a computer

Description automatically generated

Step – 11 : Cross check the columns, after renaming them

A screenshot of a computer

Description automatically generated

Step – 12 : Check which payment is mostly used

A screenshot of a computer

Description automatically generated

Step – 13 : Display maximum amounts of trips

A screenshot of a computer

Description automatically generated

Step – 14 : Display which company has more trips booked

A screenshot of a computer

Description automatically generated

Step – 15 : Now, we will create a table from the existing dataframe

A screenshot of a computer

Description automatically generated

Step – 16 : Run some more sql commands

A screenshot of a computer

Description automatically generated

Step – 17 : Plot payment\_type and its count using matplotlib

A screenshot of a computer program

Description automatically generated

A graph on a screen

Description automatically generated

Step – 18 : Convert the Trip\_Start\_Timestamp and Trip\_End\_Timestamp into timestamp datatype

A screenshot of a computer

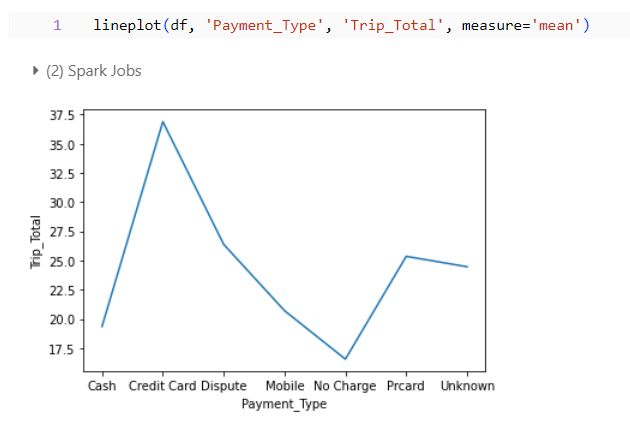
Description automatically generated

Step – 19 : Now, plot a graph b/w Trip\_Start\_Timestamp and miles

A screen shot of a computer

Description automatically generated

Step – 20 : Plot various graphs such as : histogram, lineplot and pieplot

 A graph of a line

Description automatically generated with medium confidence

A pie chart with text on it

Description automatically generated A pie chart with different colors

Description automatically generated

Step – 21 : Drop the unnecessary columns in order to run ml algorithms

A screenshot of a computer

Description automatically generated

Step – 22 : Replace all the null values with mean of the column

A screenshot of a computer code

Description automatically generated

Step – 23 : Import the necessary libraries for the linear regression algorithm

A screenshot of a computer program

Description automatically generated

Step – 24 : Apply Linear Regression algorithm for the dataframe

Rmse will measures the average differenc b/w predicted values and actual values

A screenshot of a computer

Description automatically generated

Step – 25 : Import necessary libraries for Decision Tree Algorithm



Step – 26 : Apply Decision Tree Algorithm for the dataframe

A screenshot of a computer

Description automatically generated